SELF ASSEMBLING PROTEINS FOR PRODUCING EXTENDED MATERIALS

CROSS-REFERENCE TO RELATED APPLICATIONS

This application is a continuation-in-part of application serial no. 09/564,710 filed May 3, 2000; which application claims priority to the filing date of United States Provisional Patent Application Serial No. 60/133,470 filed May 10, 1999; the disclosures of which applications are herein incorporated by reference.

ACKNOWLEDGMENT OF GOVERNMENT SUPPORT

This invention was made with Government support under Grant No. GM31299 awarded by the National Institute of Health; Grant No. MCB-0103549 from the National Science Foundation and Grant No. DE-FG03-87ER60615 awared by the Department of Energy. The Government has certain rights in this invention.

INTRODUCTION

Technical Field

The field of this invention is nanotechnology and biomaterials.

Background of the Invention

The emerging field of nanotechnology has allowed the ability to design and fabricate novel small materials with sizes or length scales in the nanometer range that can serve complex functions. These materials fall into a variety of architectural classes, such as compact clusters, hollow shells, tubes, two-dimensional layers, and three-dimensional molecular networks. These materials can subsequently be manipulated in reproducible ways to develop structures that have particular properties for novel applications. For example, such applications include the use of such structures in the development of biological coatings, such as in protein and DNA microarrays. Such nanotechnology materials have also found particular use in applications such as sensors and detectors, and

as molecular sieves for filtration.

However, in view of recent developments in the filed of nanotechnology there still remains a continued need and interest in the development of new materials and systematic methods for producing nanostructures using such materials, especially for the development of biological macromolecules for use in such applications. The present invention addresses this need.

Relevant Literature

U.S. Patents of interest include: 5,877,279; and 5,712,366. Articles of interest are: Collier, et al., Ann. Rev. Phys. Chem. (1998) 49: 371-404 (compact clusters); Rao, et al., Current Opinion in Solid State and Materials Sci. (1996) 1:279-284 and Kroto, Nature (1987) 329:529 (hollow shells); Iijima, Nature (1991)354:56-58, Ghadiri, Nature (1993)366:324-327 and Ajayan et al., Reports on Progress in Physics (1997) 60:1025-1062 (tubes); Stange, et al., Biophys. Chem. (1998) 72:73-85 (molecular networks); and Li, et al., Science (1999) 283: 1145-1147; Seeman, Trends in Biotechnology (1999)) 11:437-443 (DNA); and Chui, et al., Science (1999) 283:1148-1150 (two-dimensional layers). Also of interest are: Padilla et al., PNAS (2001) 98(5)2217-21; Dotan et al., Angew. Chem. Int. Ed. Engl. (1999) 38(16):2363-2366; Winfree et al., (1998) 394(6693):539-544; Wukowitz et al., Nature Struct. Biol. (1995) 2:1062-1067; Ringler and Schulz. Science (2003) 302(5642) 106-109.

SUMMARY OF THE INVENTION

Self-assembling fusion proteins and nucleic acids encoding the same are provided. The subject fusion proteins include a first dimer forming oligomerization domain and a second tetramer forming oligomerization domain rigidly linked to each other. Also provided are regular structures made up of a plurality of self-assembled fusion proteins of the subject invention, and methods for producing the same. The subject fusion proteins find use in the preparation of self-assembled nanostructures, e.g., two-dimensional layers and three-dimensional networks, which structures find use in a variety of different

applications.

BRIEF DESCRIPTION OF THE FIGURES

FIG 1 is a schematic illustration of a two-dimensional layer produced according to the present invention. Circles represent dimeric subunits and squares represent tetrameric subunits. The arrows indicate the 2-fold axis of the dimer units. The 4-fold axes of the tetramer come out of the page and go into the page for the black and grey squares, respectively. A single fusion protein (circled) consists of one subunit of the dimer and one subunit of the tetramer. The assembly extends indefinitely to fill the plane.

DESCRIPTION OF THE SPECIFIC EMBODIMENTS

Self-assembling fusion proteins and nucleic acids encoding the same are provided. The subject fusion proteins include a first dimer forming oligomerization domain and a second tetramer forming oligomerization domain rigidly linked to each other. Also provided are regular structures made up of a plurality of self-assembled fusion proteins of the subject invention, and methods for producing the same. The subject fusion proteins find use in the preparation of self-assembled nanostructures, e.g., two-dimensional layers, which structures find use in a variety of different applications.

Before the subject invention is described further, it is to be understood that the invention is not limited to the particular embodiments of the invention described below, as variations of the particular embodiments may be made and still fall within the scope of the appended claims. It is also to be understood that the terminology employed is for the purpose of describing particular embodiments, and is not intended to be limiting. Instead, the scope of the present invention will be established by the appended claims.

It must be noted that, as used in this specification and the appended claims, the singular forms "a," "an" and "the" include plural reference unless the context clearly dictates otherwise. Unless defined otherwise all technical and scientific terms used herein have the same meaning as commonly understood to one of ordinary skill in the art to which this invention belongs.

Where a range of values is provided, it is understood that each intervening value, to the tenth of the unit of the lower limit, unless the context clearly dictates otherwise, between the upper and lower limit of that range and any other stated or intervening value in that stated range, is encompassed within the invention. The upper and lower limits of these smaller ranges may independently be included in the smaller ranges, and such embodiments are also encompassed within the invention, subject to any specifically excluded limit in the stated range. Where the stated range includes one or both of the limits, ranges excluding either or both of those included limits are also included in the invention.

All publications mentioned herein are incorporated herein by reference for the purpose of describing and disclosing components that are described in the publications that might be used in connection with the presently described invention.

In further describing the subject invention, the subject fusion proteins, nucleic acids encoding the same and methods for producing the same are described first in greater detail, followed by a review of representative structures (as well as specifically applications in which the same find use) which may be produced from the subject fusion proteins.

FUSION PROTEINS

As summarized above, the subject invention provides fusion proteins that are capable of assembling under suitable conditions to produce regular structures, e.g., two-dimensional layers. The fusion proteins of the subject invention are characterized by having first and second oligomerization domains joined together, e.g., covalently linked or fused together, through a linking group, such as a rigid linking group. The fusion proteins of the present invention may vary in size, depending on the nature of the first and second oligomerization domains and any linker present therein. In general, the subject fusion proteins may range in length from about 50 to about 1000, including from about 75 to

about 750 such as from about 100 to about 500 aa, and have a molecular weight ranging from about 5 kDa to about 100 kDA, such as from about 8 kDa to about 80 kDa, including from about 10 kDa to about50 kDa.

The oligomerization domains may have amino acid sequences that are found in naturally occurring proteins, or may have sequences that are derivatives of sequences found in naturally occurring proteins, e.g., where the domains are mutants of naturally occurring domains (including point mutants, deletion mutants, substitution mutants, etc.). Alternatively, the oligomerization domains of the subject fusion proteins may have sequences that are not found in naturally occurring proteins, but instead are entirely synthetic. In many embodiments, however, the oligomerization domains have amino acid sequences that are found in or derived from naturally occurring proteins. By naturally occurring protein is meant a protein that occurs in nature.

The oligomerization domains or components of the subject fusion proteins are domains of stretches of amino acids that, under appropriate self-assembly conditions, such as physilogical conditions, associate with one or or more identical domains to produce a stable, multimeric structure, e.g., a dimeric structure, a tetrameric structure. While the length of a given oligomerization domain may vary, in many embodiments the length ranges from about 20 to about 500 aa, such as from about 50 to about 400 aa, including from about 80 to about 300 aa. Accordingly, in many embodiments the weight of each oligomerization domain of the subject fusion proteins may vary, but may range from about 2 to about 50 kDa, such as from about 5 to about 40 kDa, including from about 8 to about 30 kDa.

Generally, the oligomerization domains of the subject fusion proteins are either: (a) a domain or stretch of amino acids which naturally associates into a dimeric structure (e.g., it is found in a protein that associates with an identical protein to produce a dimer); or (b) a domain or stretch of amino acids which naturally associates into a tetrameric structure (e.g., it is found ina protein that associates with three identical proteins to produce a tetramer). A further characterization of the subject fusion proteins is that they include two different oligomerization domains, i.e., a first oligomerization domain and a second

oligomerization domain, where the first oligomerization domain is a domain that dimerizes with like domains, such that it is a dimerization domain, and the second oligomerization domain is a domain that associates into tetramers with three other identical domains, such that it is a tetramerization domain. Specific proteins of interest with known three-dimensional structures that naturally associate into oligomeric (e.g., dimeric or tetrameric) structures include those dimers and tetramers listed in the publicly available Protein Data Bank, described in Abola et al., Meth. Enzymol. (1997) 277:556-571, and the like.

The dimerization domains of the subject fusion proteins may have any convenient sequence of amino acids, so long as the sequence is such that the domain dimerizes with like domains, as described above. Representative dimerization domains of interest include, but are not limited to: orange carotenoid protein from A. maxima (having amino acid sequences found at Genbank accession 28373616); M1 matrix protein from influenza virus (having amino acid and encoding nucleic acid sequences found at Genbank accession no.3793307); and the like.

The tetramerization domains of the subject fusion proteins may have any convenient sequence of amino acids, so long as the sequence is such that the domain tetramerizes with like domains, as described above. Representative tetramerization domains of interest include, but are not limited to: neuraminidase from influenza virus (having amino acid and encoding nucleic acid sequences found at Genbank accession no. 37785300; E.coli fuculose aldolase (having amino acid and encoding nucleic acid sequences found at Genbank accession no. 16130707); and the like.

A feature of the subject fusion proteins is that the two or more naturally occurring protein components are joined to each other in a sufficiently rigid manner such that the orientation in space of each component relative to the other(s) in the fusion protein is relatively static and can be anticipated in advance based on the known structures of the components. Typically, the protein components of the subject fusion proteins are joined to each other through a rigid linking group that is capable of providing the requisite static orientation of the disparate components of the fusion protein. The length of the rigid

linking group may vary depending on the desired overall geometry of the fusion protein, as described below. Generally, the linking group has a length ranging from about 1.5 Å to 48 Å, such as from about 6 Å to 30 Å, including from about 6 Å to 20 Å. As such, the number of residues in the linking group generally ranges from about 1 to 35, such as from about 2 to 20, including from about 4 to 15.

Any linking group capable of providing the requisite static orientation of the disparate components of the fusion protein may be employed. As such, the linking group may hold the disparate components longitudinal to each other, such that the fusion proteins are aligned along the same axis, or the linking group may hold the components at an angle to each other, e.g., at a right angle, such that axis of the first oligomerization domain is perpendicular to the axis of the second oligomerization domain. Of particular interest in certain embodiments is the use of a linking group that includes an alpha helical structure. In other words, the linking group may include a sequence of amino acid residues that is prone to forming an alpha helix. A variety of such sequences are known and include long alpha helices found in the protein structure database such as the helix in the ribosomal protein L9 (PDB code 1div). Alternatively, it is understood that certain amino acid types tend strongly to adopt an alpha helical configuration, and the linker may be designed to contain amino acids with this tendency. In other embodiments, the linkage between the oligomerization domains may be achieved by chemical bonds between amino acid side chains

A feature of the subject fusion proteins is that they are capable of participating in a self-assembly process under suitable conditions to produce a regular, defined structure of a plurality of fusion proteins. By plurality of fusion proteins is meant at least about 2, but the number of individual fusion proteins in a particular structure may be much higher, e.g., 5, 10, 15, 20, 30, 50, 100 or more, and sometimes a very large number, particularly in essentially infinitely repeating structures. As mentioned above, the subject fusion proteins are capable of self-assembling under suitable conditions, i.e., self-assembly conditions, to produce regular structures. Suitable conditions are those conditions sufficient to provide for the self-assembly or association of the disparate fusion proteins into a regular

structure. Representative conditions under which self-assembly of the subject fusion proteins occurs are physiologic conditions or other laboratory conditions under which the individual component proteins are stable. By physiologic conditions is meant conditions found in living cell, e.g. a microbial, plant or animal cell. Typically, the conditions include an aqueous medium having a pH ranging from about 4 to 10 and usually from about 6 to 8, where the temperature ranges from about 4°C to 35°C. However, it is understood that some proteins such as those from thermophilic microorganisms are stable under very extreme conditions and that structures from such stable components may have applications under such conditions. In many embodiments, the self-assembling proteins will self-assemble with each other without the assistance or aid of accessory proteins, e.g., chaperones, etc.

The subject fusion proteins are ones that self-assemble into regular structures. By "regular structure" is meant that the structure has a defined pattern of assembly in two-dimensional or three-dimensional space that is known. In many embodiments, the subject fusion proteins self-assemble into effectively infinitely repeating regular structures, such as two-dimensional layers. The subject fusion proteins assemble into such regular structures because each oligomerization domain, e.g., dimerization domain, tetramerization domainnaturally occurring protein component, serves as an oligomerization domain which provides for the association of the fusion proteins into the regular structure. As such, the relative orientations of the disparate components of the fusion protein are selected to provide for the desired regular structure upon self-assembly under suitable conditions. Accordingly, for any given fusion protein, the relative orientation of each component thereof is chosen based on the structure into which the fusion protein is designed to self-assemble. More specifically, the geometric relationship of the symmetry elements of the oligomerization domains of the subject fusion proteins are chosen based on the desired regular structure.

The symmetry elements, i.e., symmetry axes, of a given fusion protein are configured relative to each other, i.e., have a geometry, in a manner to provide for the overall symmetry required to produce the desired structure. In many embodiments, the

geometry of symmetry elements is non-intersecting. The geometry of the symmetries of each of the components is such that they are either parallel and non-intersecting, or perpendicular and non-intersecting. In certain embodiments, the non-intersecting symmetry elements or axes form an angle that ranges from about 10 to about 90°, such as from about 25 to about 75 °, including from about 33 to about 66 °, e.g., 45°, 50°, 55°, etc. The desired geometry may be achieved either by the design features of the part of the polypeptide chain that joins the two components, or by another covalent connection between the two components, e.g., a disulfide bond or the like.

NUCLEIC ACIDS ENCODING THE FUSION PROTEINS

Also provided by the subject invention are nucleic acid compositions. By nucleic acid composition is meant a composition comprising a sequence of nucleotides having an open reading frame that encodes a fusion protein of the subject invention, as described *supra*. As such, the subject nucleic acid compositions at least include a nucleic acid sequence that encodes each of the oligomerization domains, where these sequences are generally joined by a sequence that encodes an amino acid sequence that is prone to form an alpha-helical configuration. Though the length of the subject nucleic acid compositions may vary greatly depending on the particular fusion protein that is encoded thereby, generally the subject nucleic acid compositions are at least about 60 bp long, such as at least about 150 bp long, including at least about 300 bp long, where the subject nucleic acid compositions may be as long as 3 kbp or longer, but will usually not exceed about 2 kbp in length.

The subject nucleic acid compositions may be produced by standard methods of restriction enzyme cleavage, ligation and molecular cloning. One protocol for constructing the subject nucleic acid compositions includes the following steps. First, purified nucleic acid fragments containing desired component nucleotide sequences as well as extraneous sequences are cleaved with restriction endonucleases from initial sources, e.g. animal cell, plant cell or microbial or viral genomes. Fragments containing the desired nucleotide

sequences are then separated from unwanted fragments of different size using conventional separation methods, e.g., by agarose gel electrophoresis. The desired fragments are excised from the gel and ligated together in the appropriate configuration so that a circular nucleic acid or plasmid containing the desired sequences, e.g. sequences corresponding to the various elements of the subject nucleic acid compositions, as described above, is produced. Where desired, the circular molecules so constructed are then amplified in a prokaryotic host, e.g. *E. coli*. The procedures of cleavage, plasmid construction, cell transformation and plasmid production involved in these steps are well known to one skilled in the art and the enzymes required for restriction and ligation are available commercially. (See, for example, R. Wu, Ed., Methods in Enzymology, Vol. 68, Academic Press, N.Y. (1979); T. Maniatis, E. F. Fritsch and J. Sambrook, Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y. (1982); Catalog 1982-83, New England Biolabs, Inc.; Catalog 1982-83, Bethesda Research Laboratories, Inc.

The above nucleic acid compositions find use in the preparation of the subject fusion proteins.

METHODS OF PREPARING THE SUBJECT FUSION PROTEINS

The subject fusion proteins are obtained by expressing a recombinant gene encoding the fusion proteins, such as the polynucleotide compositions described above, in a suitable host. For expression, an expression cassette may be employed. The expression vector will provide a transcriptional and translational initiation region, which may be inducible or constitutive, where the coding region is operably linked under the transcriptional control of the transcriptional initiation region, and a transcriptional and translational termination region. These control regions may be derived from a variety of sources.

Expression vectors generally have convenient restriction sites located near the promoter sequence to provide for the insertion of nucleic acid sequences encoding

heterologous proteins. A selectable marker operative in the expression host may be present. Expression cassettes may be prepared comprising a transcription initiation region, the region encoding the fusion protein, and a transcriptional termination region. After introduction of the DNA, the cells containing the construct may be selected by means of a selectable marker, the cells expanded and then used for expression. The expression cassette contained in the cell may be as part of an extrachromosomal element or integrated into the genome of the cell as a result of introducing the expression cassette into the cell. Accordingly, in many embodiemnts, the expression cassette will be maintained in the host cell and passed on to cellular progeny of the host cell.

The proteins may be expressed in prokaryotes or eukaryotes in accordance with conventional ways, depending upon the purpose for expression. For large scale production of the protein, a unicellular organism, such as *E. coli, B. subtilis, S. cerevisiae*, insect cells in combination with baculovirus vectors, or cells of a higher organism such as vertebrates, particularly mammals, *e.g.* COS 7 cells, may be used as the expression host cells. In some situations, it is desirable to express the proteins in eukaryotic cells, where the encoded protein will benefit from native folding and post-translational modifications.

Where desired, the protein may be purified following its expression to produce a purified protein comprising composition. Any convenient protein purification procedures may be employed, where suitable protein purification methodologies are described in Guide to Protein Purification, (Deuthser ed.) (Academic Press, 1990). For example, a lysate may be prepared from the original source, e.g. the expression host expressing the protein, and purified using HPLC, exclusion chromatography, gel electrophoresis, affinity chromatography, and the like.

PREPARATION OF REGULAR STRUCTURES

The subject fusion proteins find use in the production of various types of regular structures, i.e., structures of defined and predictable geometry. Specifically, the subject fusion proteins find use in the preparation of nanosized two-dimensional crystalline layers

or three-dimensional crystalline networks, where such layers or networks may in theory be of infinite size (e.g., length and width) but in many embodiments range in linear dimension from about 20 nm to about 1mm, such as from about 50 nm to about 500 um, including from about 100nm to about 200 μm,.

To prepare regular structures from the subject fusion proteins, the fusion proteins are generally combined under conditions sufficient for self-assembly of the fusion proteins into the desired regular structure to occur. Representative conditions that promote self-assembly are physiologic conditions, as mentioned above. The concentration of the fusion protein in the medium must be sufficiently high such that self-assembly into the desired structure occurs. Typically, the fusion protein concentration is at least about 0.05 mg/ml and more usually at least about 0.25 mg/ml.

In many embodiments, the structures are assembled from a plurality of identical fusion proteins, i.e., they are homogenous with respect to the fusion protein. In such embodiments, preparation of the fusion protein (e.g. expression of a nucleic acid encoding the protein) may occur in the same reaction medium as assembly of the structure, e.g., in the host cell used to express the fusion protein. Alternatively, in other embodiments, preparation of distinct fusion proteins (e.g., expression of nucleic acids encoding the proteins) may occur in separate media, prior to assembly of the heterogeneous structure.

UTILITY

The regular structures produced by the self-assembly of the subject fusion proteins find use in a variety of different applications. As mentioned above, structures can be assembled that resemble two-dimensional layers. Such ordered two-dimensional protein layers find use as biological coatings, sensors, detectors, molecular sieves, substrates for the attachment of specific chemical groups at ultra-high density, as porous materials for filtration with precise pore sizes on the mid-nanometer scale, and the like. Likewise, three-dimensional network materials may find use as sensors, porous catalysts, materials for slow release of encapsulated compounds, and the like.

KITS

Also provided are kits for use in producing the subject fusion proteins and self-assembled regular structures. The subject kits at least include a nucleic acid composition that encodes a fusion protein, where the nucleic acid is typically present on a vector. The kits may further include expression hosts suitable for expressing the subject fusion proteins. Also provided in the kits may be other reagents useful for producing the subject fusion proteins, e.g. buffers, growth mediums, enzymes, selection reagents, and the like.

In addition to above-mentioned components, the subject kits typically further include instructions for using the components of the kit to practice the subject methods. The instructions for practicing the subject methods are generally recorded on a suitable recording medium. For example, the instructions may be printed on a substrate, such as paper or plastic, etc. As such, the instructions may be present in the kits as a package insert, in the labeling of the container of the kit or components thereof (i.e., associated with the packaging or subpackaging) etc. In other embodiments, the instructions are present as an electronic storage data file present on a suitable computer readable storage medium, e.g. CD-ROM, diskette, etc. In yet other embodiments, the actual instructions are not present in the kit, but means for obtaining the instructions from a remote source, e.g. via the internet, are provided. An example of this embodiment is a kit that includes a web address where the instructions can be viewed and/or from which the instructions can be downloaded. As with the instructions, this means for obtaining the instructions is recorded on a suitable substrate.

The following examples are offered by way of illustration and not by way of limitation.

EXPERIMENTAL

In order to demonstrate a two-dimensional protein layer, self-assembling into a square, repeating pattern, we fused a cyclic tetrameric protein, fuculose aldolase from E. coli, to a dimeric protein, M1 matrix protein from influenza virus. These two oligomerization domains were connected in the larger fusion protein by a short flexible linker. Specifically, residues 1-206 of fuculose aldolase were fused to residues 1-187 of the M1 matrix protein via a 5-residue linker with amino acid sequence DPVPV. This fusion resulted in a 398 residue fusion protein with an N-terminal fuculose aldolase domain and a C-terminal M1 matrix protein domain, having a molecular weight of 41 kDa. The geometrically rigid connection between the two domains was achieved by way of a covalent bond between individual cysteine residues in the two oligomerization domains. These cysteine residues were genetically engineered into the two domains in advance, specifically at residue 86 of fuculase aldolase and residue 34 of M1 matrix protein. According to the geometry of the design, with the symmetry axes of the two oligomerization domains being non-intersecting but forming an angle very nearly equal to 90 degrees, the fusion proteins are designed to assemble into an extended two-dimensional layer with a symmetry that would be described by the notation $p42_12$.

A multistep PCR protocol was used to create the construct described above from the mutated dimer and tetramer clones. This construct was ligated into a pET-21b vector between the NdeI and EcoRI sites, adding a histidine tag to the N-terminus, where it is not expected to interfere with the intramolecular disulfide formation or subsequent layer assembly. This vector was transformed into *E. coli* BL21-DE3 cells. The cells were grown at 37° C and induced with IPTG at an optical density of 0.6. The protein is produced in milligram quantities, but is present inside the host cells in inclusion bodies. Therefore, the protein was purified from inclusion bodies in the presense of dithiothreitol and urea.

The resultant purified protein self assembles into a two-dimensional protein layer, having a square, repeating pattern.

It is evident from the above results and discussion the subject invention provides powerful tools and methodologies for producing ordered structures from self-assembling fusion proteins. The fusion proteins of the subject invention can be readily produced and

then self-assembled into a variety of different structures which find use in a plurality of different applications. As such, the subject invention represents a significant contribution to the field.

All publications and patent applications cited in this specification are herein incorporated by reference as if each individual publication or patent application were specifically and individually indicated to be incorporated by reference. The citation of any publication is for its disclosure prior to the filing date and should not be construed as an admission that the present invention is not entitled to antedate such publication by virtue of prior invention.

Although the foregoing invention has been described in some detail by way of illustration and example for purposes of clarity of understanding, it is readily apparent to those of ordinary skill in the art in light of the teachings of this invention that certain changes and modifications may be made thereto without departing from the spirit or scope of the appended claims.